DARPA CS Study Panel 2007: Final Report

Noah Smith, Carnegie Mellon University, nasmith@cs.cmu.edu

1. Programmatic

I have attended the four Computer Science Study Panel (CSSP) sessions, in April, June, July, and October 2007. I found the sessions to be highly interesting, both as a citizen and as a scientist. Because my research deals with information – in particular, automated processing of language data like text and speech – I saw the greatest connection with my research in the final session when we interact with the intelligence community. I also visited the National Security Agency in August. I gave an overview of my research there and spoke informally with various NSA personnel to learn more about NSA's interests in natural language technology, which I knew prior to this program to be quite strong.

The impact of this program on my research vision was huge. The three key lessons for me are as follows. First, the DOD has more non-technical challenges than it has technical ones; some of these can be addressed by improving basic technical competence and understanding of technology across all levels of the DOD. This probably requires a fundamental change in the way computing (and engineering in general) is presented in college curricula and before. As in many other US organizations, the average DOD employee simply does not understand computers well enough to know how to critically appraise the tools him to help him do his work, or contemplate how they might be made more sophisticated in non-trivial ways. Second, the "information overload" problem is not a problem of filtering; it is a problem of visualization and presentation. "More" data and information are only overwhelming if they are presented badly. Third, intelligent computer systems need to be placed in the hands of DOD people much earlier in the research process; much of what the field works on today will assuredly hold no benefit for the DOD, and will be evaluated by DOD persons much too late to have an effect. This interaction need not be via demonstrations and requirements statements; it would preferably be through natural conversations with the men and women who might use the end-products of academic research, even if the researchers themselves are not creating those end-products.

2. Technical

My technical interest is in natural language understanding by computer programs. "Understanding" is an ill-defined goal; normally researchers in my field focus on specific intelligent behaviors (such as summarizing documents or translating sentences between languages), more natural user interfaces (such as speech recognition and synthesis), or deep "understanding" within narrow domains. My particular interest is in core natural language processing tasks (such as syntactic parsing) and learning algorithms that can build these capabilities out of data.

The third quarter of my partieipation in the CS Study Group was foeused heavily on the submission of a Phase 2 proposal. An abstract of the proposal follows.

We propose to develop RAVINE, a tool for Recombination, Aggregation, and Visualization of Information in Newsworthy Expressions. RAVINE will automatieally produee metadata annotating freely available intelligenee (specifically, news articles) to show statements attributed to various individuals over time, by different reporters.

RAVINE will perform semantic proeessing on these attributions to assist the intelligence analyst in tracking multiple potentially conflicting positions over time. Further, RAVINE will provide an interface to permit querying and browsing large eolleetions of news stories based on this metadata, aiding human analysts who wish to understand not only what has been reported in the news, but also variation in news aeeounts around the world. RAVINE will bring state-of-the-art statistieal semantic modeling to bear on the defense analyst's ehallenge of aggregating information from diverse intelligence sources. Importantly, RAVINE is intended for real-world use by the end of the projeet, requiring robustness and the ability to broadly eover many topie domains. Statistieal natural language processing (NLP) technology is a powerful paradigm for building robust software for annotating text data with linguistie analysis (metadata). It enables designers to train systems from real world data instead of designing them by hand, resulting in systems that behave appropriately in the faee of new or noisy inputs. The development of RAVINE will require solutions to several important research problems. Techniques for identifying relationships between news-reported natural language statements made by different people, at different times, and in different eontexts will be essential. By automatieally classifying these relationships, we ean provide a novel visualization of reported news with respeet to a given speaker (or organization), event, topie, or point in time. The relationships we aim to model include "eehoes" (repetition of the same idea), topieal eonnection (statements relevant to eaeh other), eonsistent statements, agreeing statements, disagreeing statements, and statements that show a shift in attitude or belief over time. We note that earlier research has studied problems of paraphrase (McKeown, 1979) and textual entailment (Giampieeolo et al., 2007). In order to address this diverse set of semantic phenomena, we will make use of semantic parsing. Semantic parsing is any automated method that eonstruets a meaning representation—a type of metadata—from a piece of text. There are many different target meaning representations to ehoose from. Rather than starting from a particular meaning representation meant to serve all purposes, we take the view that different features of meaning are required for different tasks. As we move from simpler relationships (such as eehoes) to deeper ones (sueh as agreement), "deeper" semantic representations will be required. A key element of our research is the

design and large-scale data-driven development of semantic parsers for the particular problems identified. Unlike prior work on semantic parsing (Thompson, Mooney, and Tang, 1997; Zettlemoyer and Collins, 2005), which started with extremely narrow text domains and fixed first-order logic meaning representations, we will start with broad-domain newstext and develop minimally expensive meaning representations that suit our task.

The main product of this research project will be a Web interface that permits browsing of existing news streams in an entirely new way. The intelligence or defense analyst who uses our system will see the news from many sources recombined into a single coherent view that gives the aggregate of information conveyed. By focusing on reported statements and the individuals that made them, our system will provide a natural visualization of a news story in different informational contexts. Attributed statements carry a surprisingly large amount of information in a news story, as can be seen in figure 1. This work goes far beyond existing news aggregation techniques in that we will perform natural language analysis of the content of the news rather than relating stories only at the document level. The combined development of the web tool and the contributions made to language processing research areas discussed above make RAVINE an important step toward large-scale language systems and deeper language understanding.

Our most promising transition partners for Phase 3 are the intelligence agencies: the NSA, the CIA, and the DIA.

3. Funding

Incurred expenses:

$29,089	salary (PI, June-August 2007)
$6,739	benefits (PI, June-August 2007)
$21,585	salary and tuition (graduate assistant, Sept.-Dec. 2007)
$4,001	operating expenses (computer support)
$7,656	domestic travel (PI – panel sessions 1-4 and HLT-NAACL 2007, NSA visit August 2007, 2 Ph.D. students to HLT-NAACL 2007)
$2,861	foreign travel (PI – ACL 2007, June-July 2007)
$71,932	**total direct costs**
$28,260	indirect costs
$100,192	**total**

I have incurred expenses very slightly exceeding the total budget; the difference will be covered by my discretionary funds at Carnegie Mellon.